ESTIMATIONS BASED ON CROSS-ENTROPY

VÁVRA FRANTIŠEK, NOVÝ PAVEL, NEUMANOVÁ MARTINA, VOKÁČOVÁ KATEŘINA

Abstract. The minimization of the parametr v in the equivalent $-\frac{1}{n}\sum_{i=1}^{n}\lg e(x_i,v)$ in the cross-entropy $H(X;e)=-\sum_{x\in X}p(x)\lg(e(x,v))$ where $p(v)$ is a real probability distribution an $e(x,v)$ a parametrical model with the parametr v leads to estimations with maximal likelihood. (It is proved in [1,2].) This work concerns about different kinds of estimations with defined constraints and investigates the discrete distributions with smoothness conditions.

1. The smoothness of descrete destributions

The natural smoothness measure of discrete distribution $p(x);\ x\in X;$ $\sum_{x\in X}p(x)=1$ is its entropy

$$H(X)=-\sum_{x\in X}p(x)\log_2 p(x)=\tfrac{1}{\lg 2}\left(-\sum_{x\in X}p(x)\lg p(x)\right).$$

The entropy is expressed on the base of natural logarithm $H(X)=-\sum_{x\in X}p(x)\lg p(x)$ which differs from the general form in multiplicative constant $\frac{1}{\lg 2}$. The entropy by itself is the impropriate constraint for optimization in optimizing problems. Therefore a new entropy measure is modified as $N(X)=\sum_{x\in X}\lg p(x)$. If an alphabet of elementary events is finite and $\forall x\in X, p(x)>0$, the extreme of $N(X)$ are found by apparatus:

$$Q(p(x);x\in X;\lambda)=\sum_{x\in X}\lg p(x)+\lambda\left(\sum_{x\in X}p(x)-1\right)$$

$$\tfrac{\partial Q}{\partial p(x)}=\tfrac{1}{p(x)}+\lambda\quad\text{and}\quad\tfrac{\partial^2 Q}{\partial p(x_1)\partial p(x_2)}=-\tfrac{1}{p^2(x_1)}\Leftrightarrow x=x_1=x_2;\ \text{else}=0$$

then the saddle point $\Leftrightarrow\frac{\partial Q}{\partial p(x)}=0$ will be maximum. $\frac{1}{p(x)}+\lambda=0\Rightarrow p(x)=\frac{1}{|X|}$ and $N\left(\frac{1}{|X|}\right)=-|X|\lg(|X|)$. On the base of the maximum which we get for the uniform distribution is defined a standardised rate of the certainty:

$$U(X)=-\tfrac{1}{|X|\lg(|X|)}\sum_{x\in X}\lg p(x)=K\sum_{x\in X}\lg p(x);\quad K=-\tfrac{1}{|X|\lg(|X|)}.$$

The reason why is this expression called the rate of the certainty will be shown on the Bernoulli distribution $X=\{x_1,x_2\};\ p(x_1)=p,\ p(x_2)=1-p$. The measure

of the certainty gets a minimum (equal to one) for the uniform distribution and it is increasing with the growing deviation from the uniform distribution.

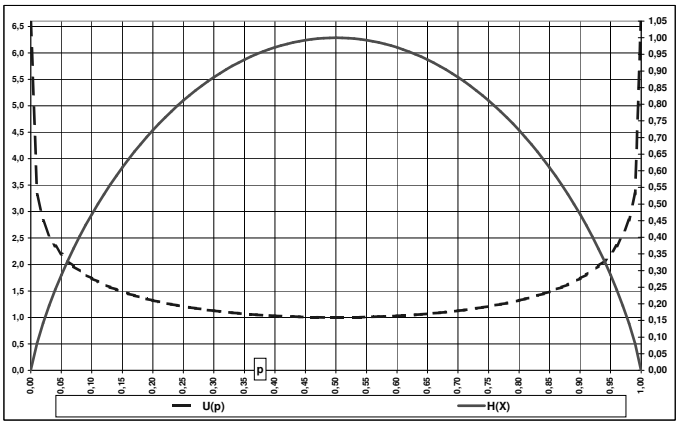

FIGURE 1. Comparison of rate of entropy and rate of certainty for Bernoulli distribution

2. The Probability Estimations of Finite Discrete Distributions with a Smoothness Condition

The probability estimations of the finite discrete distribution from observed values will be based on the following method: Let's have n observed values $(x_1, x_2, \ldots, x_n; x_i \in X)$ for a phenomenon with possible results from the alphabet X specified with model probabilities $p(x); x \in X$. Then the sample entropy is $H_n(X) = -\frac{1}{n}\sum_{i=1}^{n} \lg p(x_i)$. Optimal estimations of $p(x); x \in X$ in accordance with the smoothness condition $\sum_{x\in X} \lg p(x) = h$ are the solutions of an optimization problem:

$$Q(p, \lambda, \psi) = -\tfrac{1}{n}\sum_{i=1}^{n} \lg p(x_i) + \lambda\left(\sum_{x\in X} p(x) - 1\right) + \psi\left(\sum_{x\in X} \lg p(x) - h\right) =$$

$$= -\tfrac{1}{n}\sum_{x\in X} n(x) \lg p(x) + \lambda\left(\sum_{x\in X} p(x) - 1\right) + \psi\left(\sum_{x\in X} \lg p(x) - h\right) \overrightarrow{{}_{p(x)}}\ min$$

where $n(x)$ is the number of observations of the value $x \in X$ in the sample $(x_1, x_2, \ldots, x_n; x_i \in X)$. By the standard way we get:

$$\frac{\partial Q}{\partial p(x)} = -\left(\frac{n(x)}{n} - \psi K\right)\frac{1}{p(x)} + \lambda$$

$$\frac{\partial^2 Q}{\partial p(x_1)\partial p(x_2)} = \left(\frac{n(x)}{n} - \psi K\right)\frac{1}{p^2(x_1)} \Leftrightarrow x = x_1 = x_2;\ \text{else} = 0$$

$\Rightarrow$ if $\forall x \in X; \left(\frac{n(x)}{n} - \psi K\right) > 0$, then the saddle point represents the minimum. In a case that $\forall x \in X; \left(\frac{n(x)}{n} - \psi K\right) < 0$ the saddle point represents the maximum. For the saddle point $\frac{\partial Q}{\partial p(x)} = 0$ we get

$$p(x) = \frac{1}{\lambda}\frac{n(x)}{n} + \frac{1}{\lambda}\frac{\psi}{|X|\lg(|X|)}$$

and from the condition $\frac{\partial Q}{\partial \lambda} = 0 \Leftrightarrow 1 = \sum_{x \in X} p(x)$ is $\lambda = 1 + \frac{\psi}{\lg(|X|}$. Then

$$p(x) = \frac{1}{\lambda}\frac{n(x)}{n} + \frac{1}{\lambda}\frac{\lambda - 1}{|X|}$$

and after the substitution $\frac{1}{\lambda} = 1 - \alpha$ we obtain:

$$p(x) = (1-\alpha)\frac{n(x)}{n} + \alpha\frac{1}{|X|}.$$

However this expression represents the mixture of observed relative frequencies $(1-\alpha)$ and the uniform distribution (α). The value of α is possible to compute by numerical solving of the equation:

$$\frac{\partial Q}{\partial \psi} = 0 \Leftrightarrow h = K \sum_{x \in X} \lg p(x) = K \sum_{x \in X} \lg\left((1-\alpha)\frac{n(x)}{n} + \alpha\frac{1}{|X|}\right),$$

if such solution exists and $0 \leq \alpha \leq 1$. After substitution in the existence condition of minimum $\left(\forall \in X; \left(\frac{n(x)}{n} - \psi K\right) > 0\right)$, we get $\forall \in X$:

$$\left(\frac{n(x)}{n} - \left(\frac{-1}{|X|lg|X|} lg|X| \frac{\alpha}{1-\alpha}\right)\right) = \frac{n(x)}{n} + \frac{1}{|X|}\frac{\alpha}{1-\alpha} > 0 \Leftrightarrow (n(x) > 0) \vee (\alpha > 0).$$

The condition of non-negativity is valid for almost all cases.

An existence of the solution is dependent on a reality–reflection of the value h. The following analysis could be used for proving that the value h is realistic and for choosing "starting values" of a numerical calculation as well:

$$h(\alpha) = K \sum_{x \in X} \lg p(x) = \frac{-1}{|X|lg|X|} \sum_{x \in X} lg\left((1-\alpha)\frac{n(x)}{n} + \alpha\frac{1}{|X|}\right) =$$

$$= \frac{1}{|X|lg|X|} \sum_{x \in X} lg\left(\frac{1}{\frac{n(x)}{n} + \alpha\left(\frac{1}{|X|} - \frac{n(x)}{n}\right)}\right)$$

and

$$\frac{dh}{d\alpha} = \frac{1}{|X|\lg|X|} \sum_{x \in X} \left(\frac{n(x)}{n} + \alpha\left(\frac{1}{|X|} - \frac{n(x)}{n}\right)\right)\left(\frac{1}{|X|} - \frac{n(x)}{n}\right) =$$

$$= \frac{1}{|X|\lg|X|}\left(\frac{1-\alpha}{|X|} - (1-\alpha)\sum_{x \in X}\left(\frac{n(x)}{n}\right)^2\right)$$

furthermore is

$$1 = \sum_{x \in X}\frac{n(x)}{n} \Rightarrow 1 = \left(\sum_{x \in X}\frac{n(x)}{n}\right)^2 =$$

$$= \sum_{x \in X}\left(\frac{n(x)}{n}\right)^2 + \sum_{x,y \in X; x \neq y}\left(\frac{n(x)}{n}\right)\left(\frac{n(y)}{n}\right) \Rightarrow \sum_{x \in X}\left(\frac{(n(x)}{n}\right)^2 < 1.$$

Because:

$$\frac{dh}{d\alpha} = \frac{1}{|X|\lg|X|}\left(\frac{1-\alpha}{|X|} - (1-\alpha)\sum_{x\in X}\left(\frac{(n(x)}{n}\right)^2\right) < \frac{1}{|X|\lg|X|}\left(\frac{1-\alpha}{|X|} - (1-\alpha)\right) < 0$$

and so $h(\alpha)$ is decreasing function for $0 \leq \alpha \leq 1$ which has the maximum in $\alpha = 0$ (even in an improper point - in the case that $\exists x \in X; n(x) = 0$) and the minimum in $\alpha = 1; h(1) = 1$ The demonstration of such function for observed frequencies is:

Points	**Observed frequencies**	**Points**	**Observed frequencies**
0	1	13	19
1	3	14	23
2	5	15	29
3	7	16	18
4	7	17	15
5	18	18	18
6	16	19	10
7	13	20	9
8	25	21	13
9	19	22	3
10	19	23	3
11	31	24	0
12	22	25	1
		Total	347

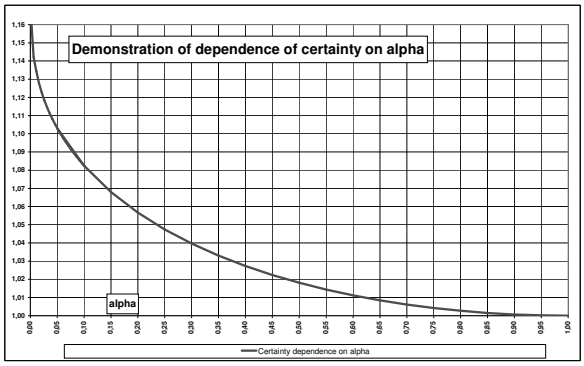

FIGURE 2. The demonstration of the dependence of the certainty on the parameter α

The mentioned single-valued relation between α and h can cause modification the smoothness of an estimated probability distribution directly by α. In this way α represents in more suitable way the smoothness. $\alpha = 1$ represents "absolute smoothness" in the case that model of the probability distribution is uniform distribution. On the other hand $\alpha = 0$ represents the not smoothed estimation –

it means that the estimation is represented only by a relative frequency of observed values. The problem and $\alpha(n)$ asymptotical behaviour of compared to amount of observed data was studied in [3].

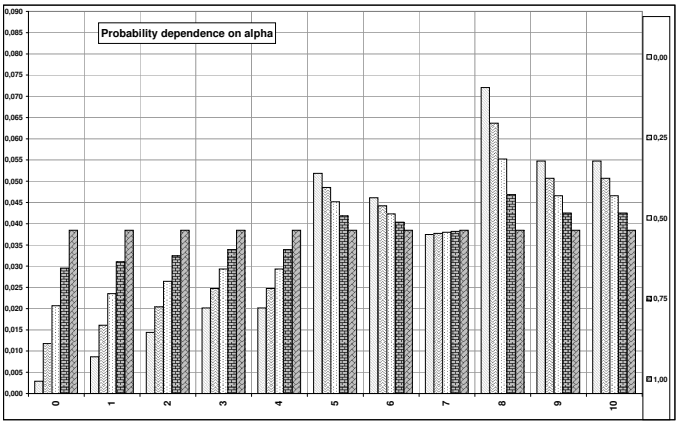

FIGURE 3. The dependence of probability estimations on the parameter α demonstrated by particular observations

3. The Probability Estimations of Finite Discrete Distributions Based on Moment Condition

The classical extremal problems of probability theory are constructions of distributions with some pre-defined moments and maximal entropy [4]. Similar problems could be solved by methods of statistics as well. Let's have a discrete random variable with a value from a set $C = \{c_1, ..., c_m\}; c_i \in R_1$ and its observed values $(x_1, x_2, ..., x_n, x_i \in C)$. Its model probabilities $p(x); x \in X$ based on the condition of some pre-assigned moment $m_k = \sum\limits_{x \in C} x^k p(x)$ are computed by solving of the optimization problem:

$$Q(p, \lambda, \psi) = -\frac{1}{n} \sum_{i=1}^{n} \lg p(x_i) + \lambda \left(\sum_{x \in C} p(x) - 1 \right) + \psi \left(\sum_{x \in C} x^k p(x) - m_k \right) =$$

$$= -\frac{1}{n} \sum_{x \in C} n(x) \lg p(x) + \lambda \left(\sum_{x \in C} p(x) - 1 \right) + \psi \left(\sum_{x \in C} x^k p(x) - m_k \right) \xrightarrow[p(x)]{} min.$$

By standard procedure we get:

$$\frac{\partial Q}{\partial p(x)} = -\left(\frac{n(x)}{n} \right) \frac{1}{p(x)} + \lambda + \psi x^k,$$

$$\frac{\partial^2 Q}{\partial p(x_1) \partial p(x_2)} = \left(\frac{n(x)}{n} \right) \frac{1}{p^2(x_1)} \Leftrightarrow x = x_1 = x_2; \text{else} = 0 \Rightarrow$$

a solution of the normal equations is the minimum. For the saddle point $\frac{\partial Q}{\partial p(x)} = 0$ we get:

$$p(x) = \frac{n(x)}{n} \frac{1}{\lambda + \psi x^k}$$

and λ, ψ are numerical solutions of following equations:

$$1 = \sum_{x \in C} p(x) = \frac{1}{n} \sum_{x \in C} \frac{n(x)}{\lambda + \psi x^k}$$

$$m_k = \sum_{x \in C} x^k p(x) = \frac{1}{n} \sum_{x \in C} \frac{x^k n(x)}{\lambda + \psi x^k}.$$

Further, there are defined family of functions with the aim to qualify a value of the parameter m_k like realistic and to compare possibilities of numerical solutions of both equations:

$$f_i(\lambda, \psi) = \sum_{x \in C} x^i p(x) = \frac{1}{n} \sum_{x \in C} \frac{x^i n(x)}{\lambda + \psi x^k}; i = 0, 1, ...,$$

represent the right sides of the equations. Their partial derivative are:

$$\frac{\partial}{\partial \lambda} f_i(\lambda, \psi) = -\frac{1}{n} \sum_{x \in C} \frac{x^i n(x)}{(\lambda + \psi x^k)^2}; i = 0, 1, ...$$

$$\frac{\partial}{\partial \psi} f_i(\lambda, \psi) = -\frac{1}{n} \sum_{x \in C} \frac{x^{(i+k)} n(x)}{(\lambda + \psi x^k)^2}; i = 0, 1, ...$$

From partial derivatives is evident, that functions $f_i(\lambda, \psi)$ are decreasing for all values of both parameters if the random variable is positive. This may simplify the choice of the solution method. As well the problem of mentioned optimization method is not considering the non-negativity condition of the probability:

$$\forall x \in C; p(x) \geq 0 \Leftrightarrow \forall x \in C; \frac{1}{\lambda + \psi x^k} > 0 \Leftrightarrow \forall x \in C; [\lambda + \psi x^k > 0].$$

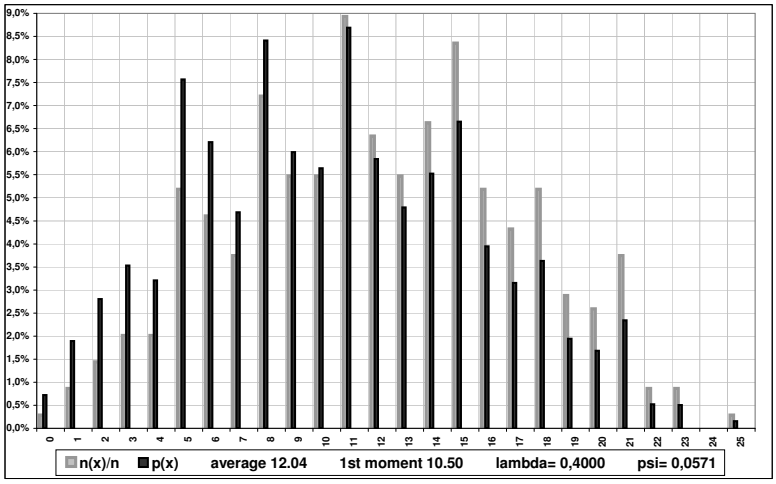

FIGURE 4. The example of probability estimation where the artificial mean value $m_1 = 10.5$ and observed average is 12.04

References

[1] F.Vávra, P.Nový; Information and disinformation. Seminar of Applied mathematics. Department of Applied Mathematics. Masaryk University - Faculty of Science, Brno 13.4.2004. In Czech
[2] F.Vávra, M.Kotlíková, H.Mašková, A.Netrvalová, P.Nový, D.Spíralová, D.Zmrhal; Information and disinformation a statistical view. Robust 2004, Třešť 2004, In Czech.
[3] F.Vávra, P.Nový, L.Reismüllerová, K.Vokáčová, M.Neumanová; Discrete Kernels. AUSTRIAN JOURNAL OF STATISTICS, Vol. 35 (2006), Number 2&3, 365–370
[4] T.M.Cover, J.A.Thomas; Elements of Information Theory. Wiley, New York 1991.
[5] L.Devroye, L.Gyorfi; Nonparametric Density Estimation: The L1 View, John Wiley, New York, 1985.

Department of Computer Science and Engineering, Faculty of Applied Science, University of West Bohemia, Univerzitni 22, 30614 Pilsen, Czech Republic
E-mail address: vavra@kiv.zcu.cz, novyp@kiv.zcu.cz, vokac@kiv.zcu.cz, mneumano@kiv.zcu.cz